Capstone Project -3

Bank Marketing Effectiveness Prediction

Done By :-

Vinit Ladse
Gaurav Bhakte
Pratiksha Kharode

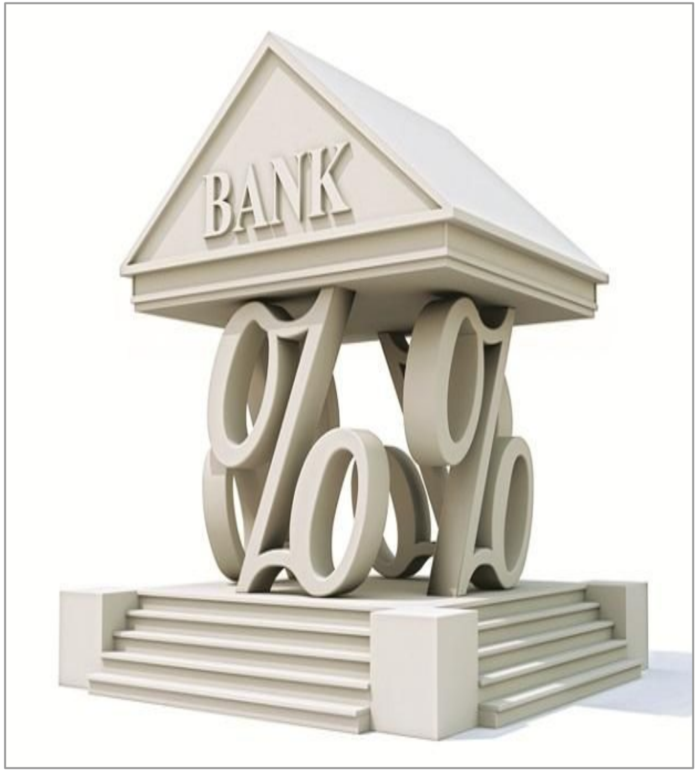
Content

- Problem Statement
- Data Summary
- Exploratory Data Analysis
- Model Implementation
- Evaluation Metrics
- Challenges
- Conclusion

Problem Statement

The data is related with direct marketing campaigns of a Portuguese banking institution.The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The classification goal is to predict if the client will subscribe a term deposit (variable 'y').



Data Summary

The Dataset contains 17 Features with 45211 observation.

Categorical Features

- Marital - (Married , Single , Divorced)
- Job - (Management,BlueCollar,retired etc)
- Contact - (Telephone,Cellular,Unknown)
- Education - (Primary,Secondary,Tertiary)
- Month - (Jan,Feb,Mar,Apr,May etc)
- Poutcome - (Success,Failure,Other,Unknown)
- Housing - (Yes/No)
- Loan - (Yes/No)
- Default - (Yes/No)

Desired target

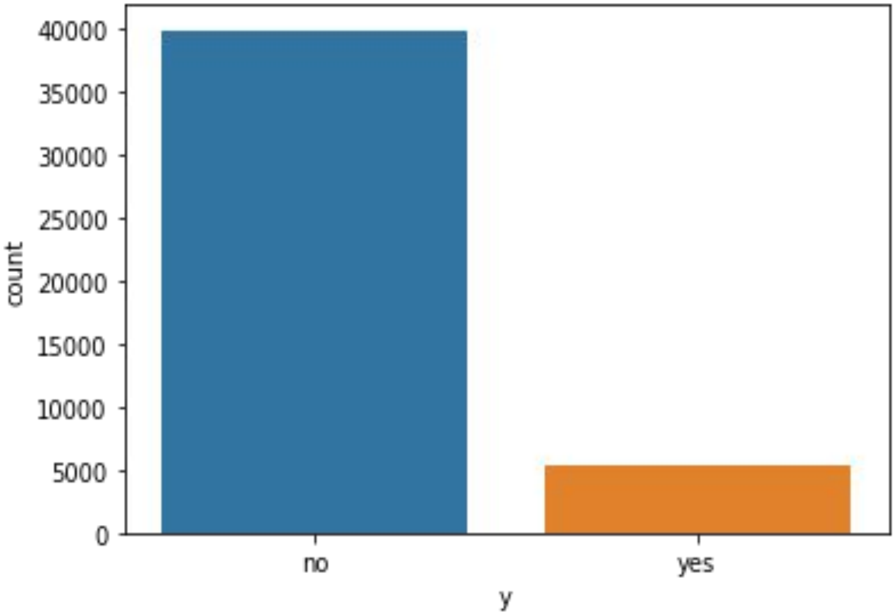
- y - has the client subscribed a term deposit? (binary: 'yes','no')

Numerical Features

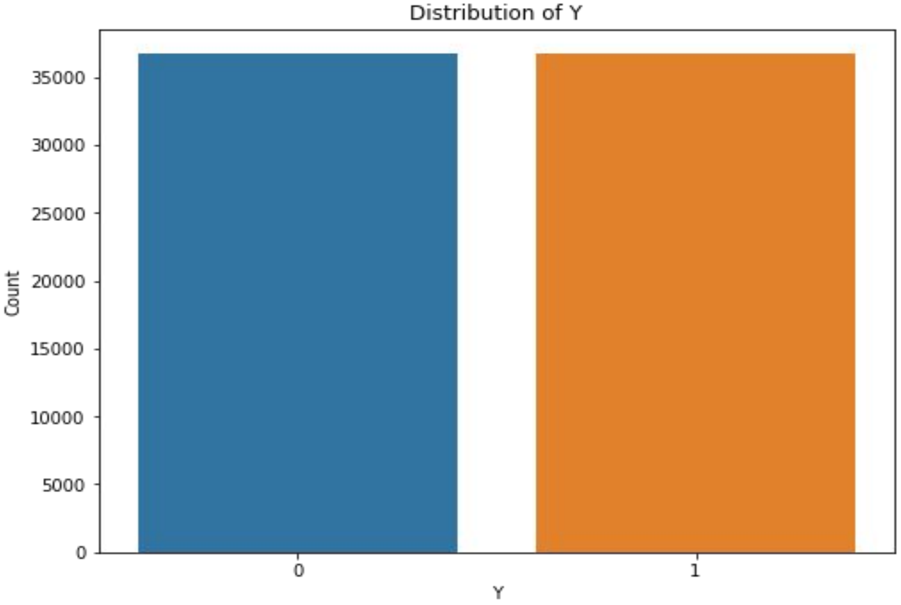
- Age
- Balance
- Day
- Duration
- Campaign
- Pdays
- Previous

Exploratory Data Analysis (Target)

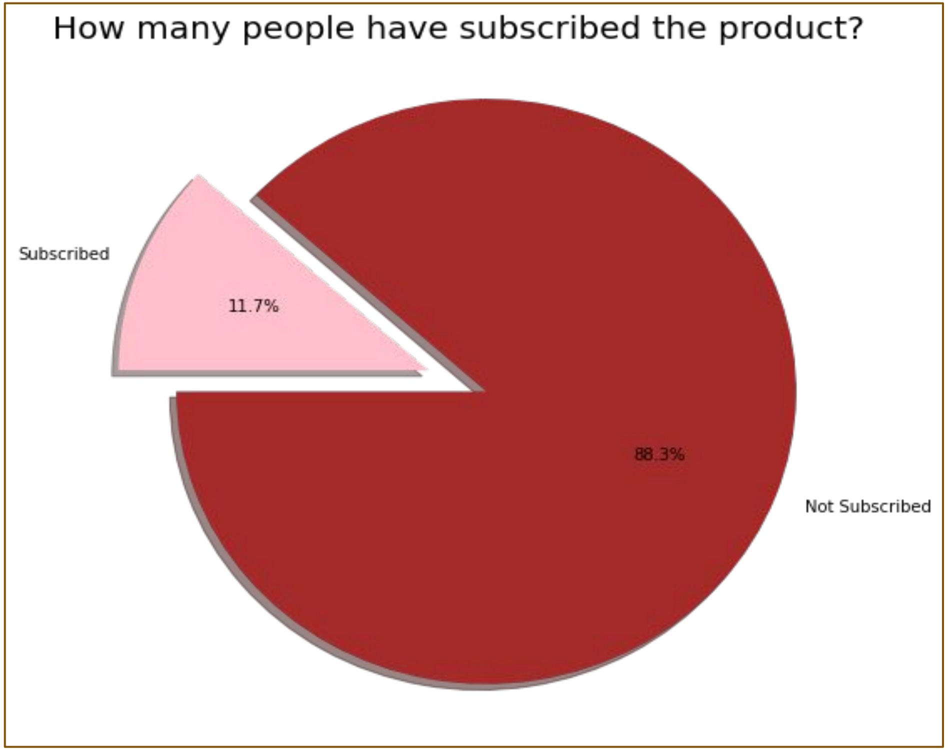
:- Before



:- After



EDA(Continued...)

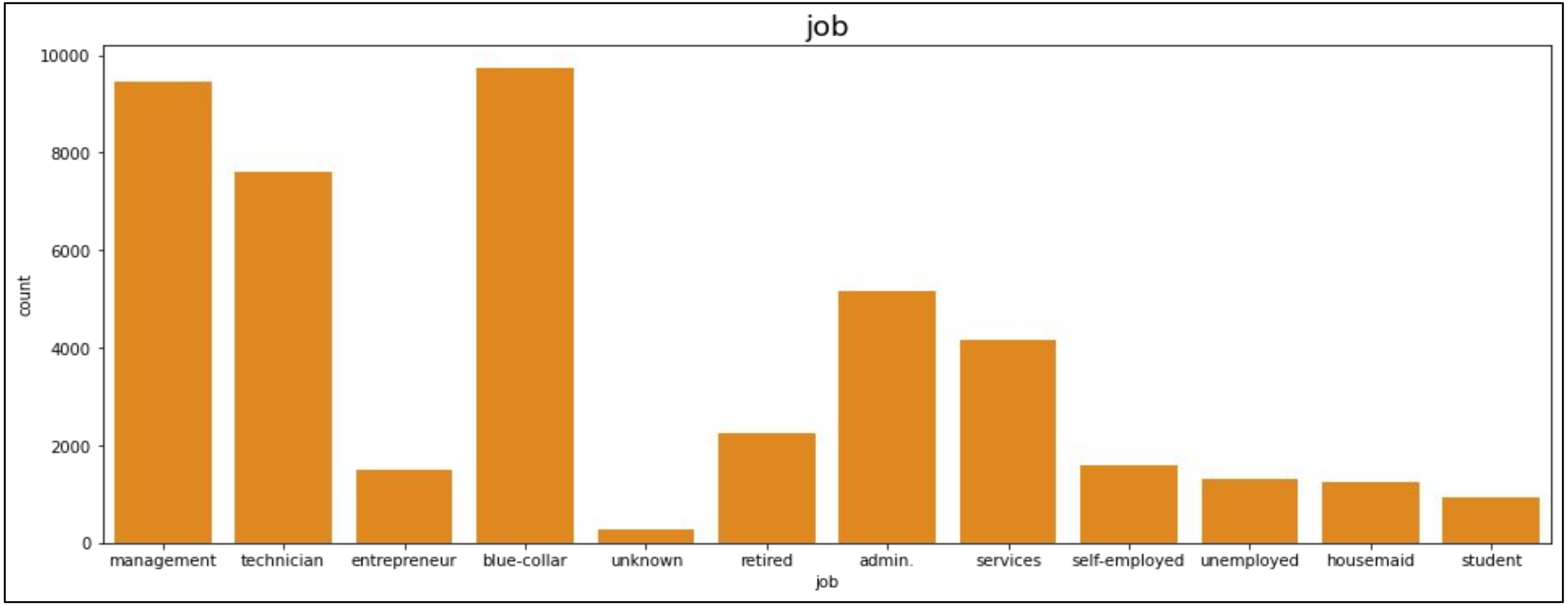


- From this data we can see that 88% customers did not subscribed for Term deposit

- We can say that the percentage of people subscribing to the term deposit is quite low.

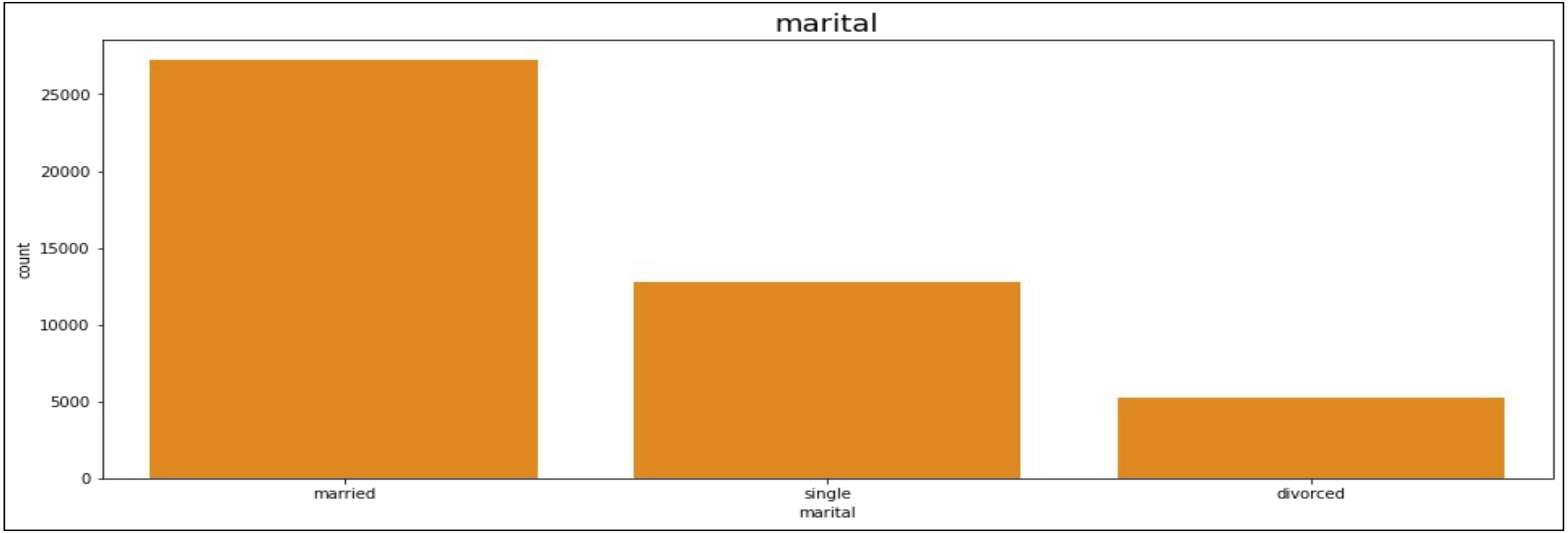
EDA(Continued...)

Categorical Features Exploration :-



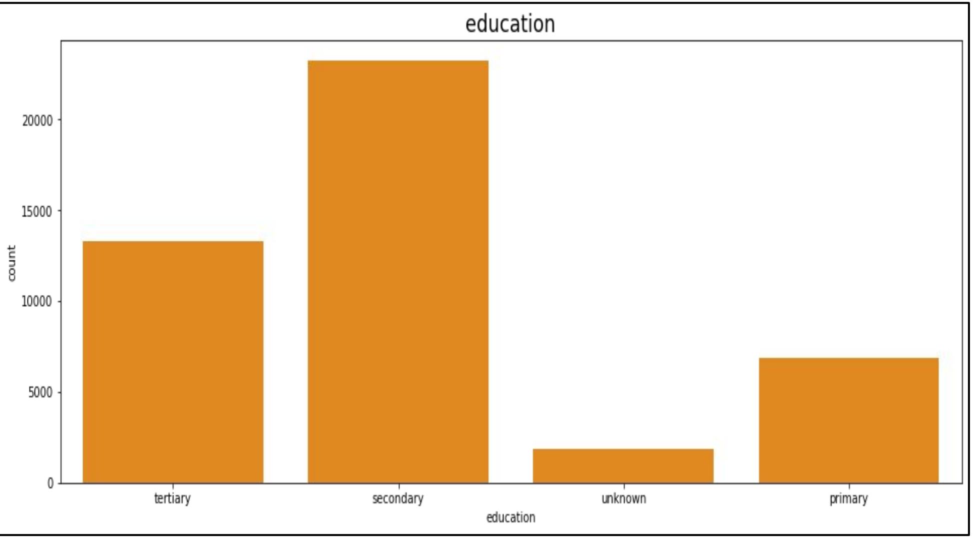
- In this plot there are 12 different job profile. Top 3 job profiles are "management", "blue-collar" and "technician“,which contains 60% of the total records.
- People with blue-collar & management jobs have subscribed more for the deposits.

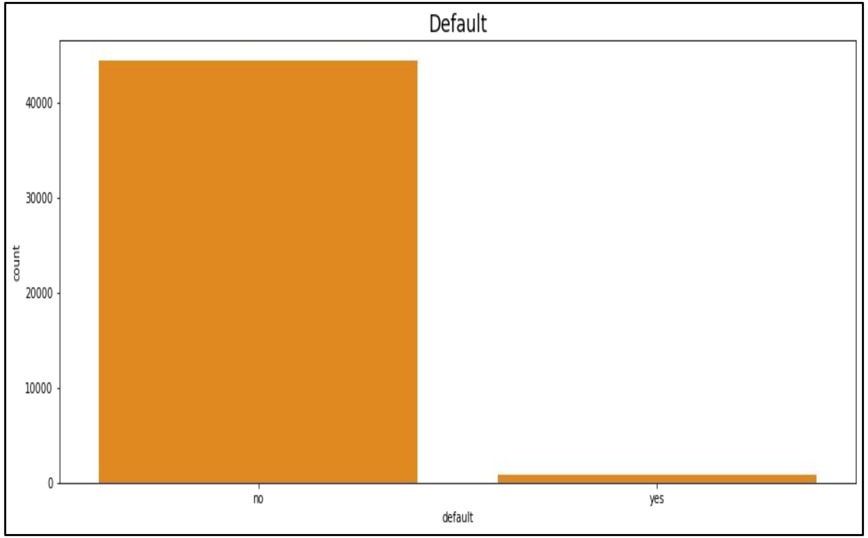
EDA(Continued...)



- Around 60% of our client base is married, 25% is single & 12% are divorced.
- Client who married are high in records.
- People who are married have subscribed for deposits more than people with any other marital status.

EDA(Continued...)

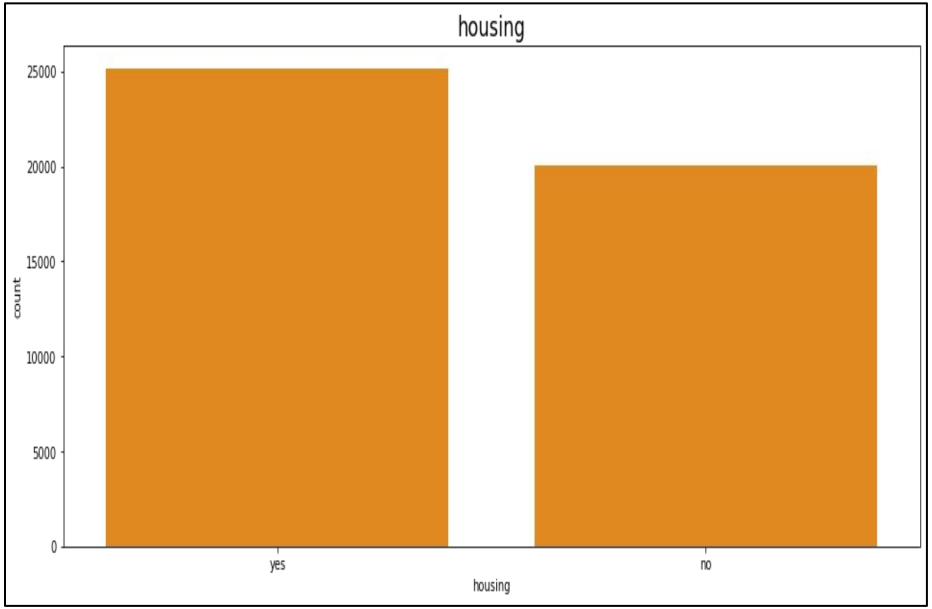


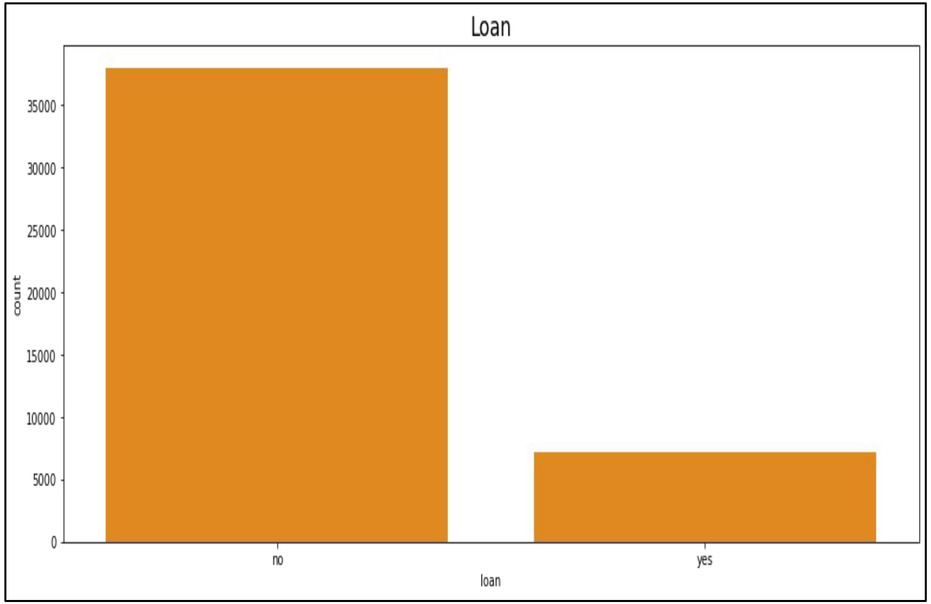


- Client whose education background is secondary are in high numbers.
- People with Secondary education qualification are the most who have subscribed for the deposits.

- Default feature seems to be does not play important role.
- People with default status as ‘no’ are the most ones who have not subscribed for bank deposits.

EDA(Continued...)

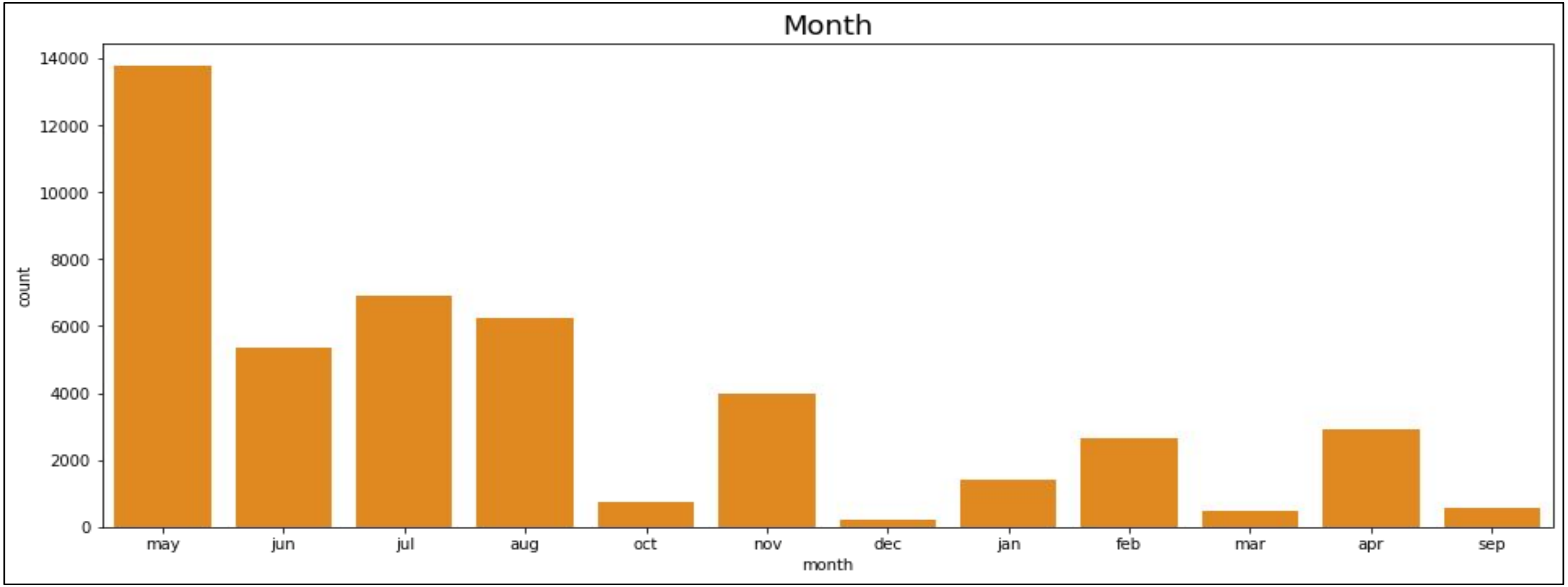




- People with housing loan are the most ones who have been contacted by the bank, followed by people with no housing loan.
- Most of the client has taken the housing loan.

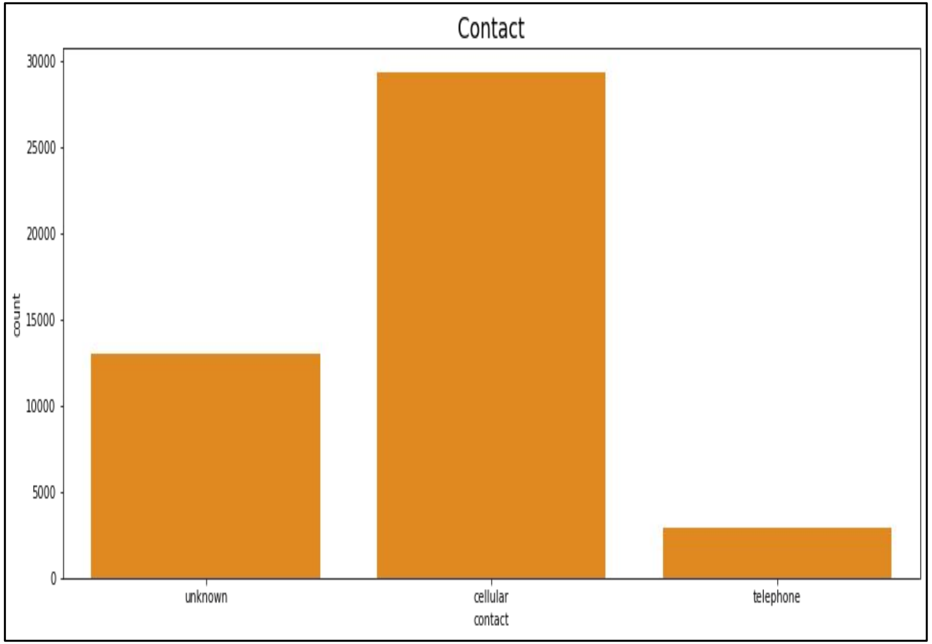
- People with no personal loan are the most ones who have been contacted by the bank for the deposits.
- People with no personal loan are the most ones who have not subscribed and are also the most ones who have subscribed for the deposits.

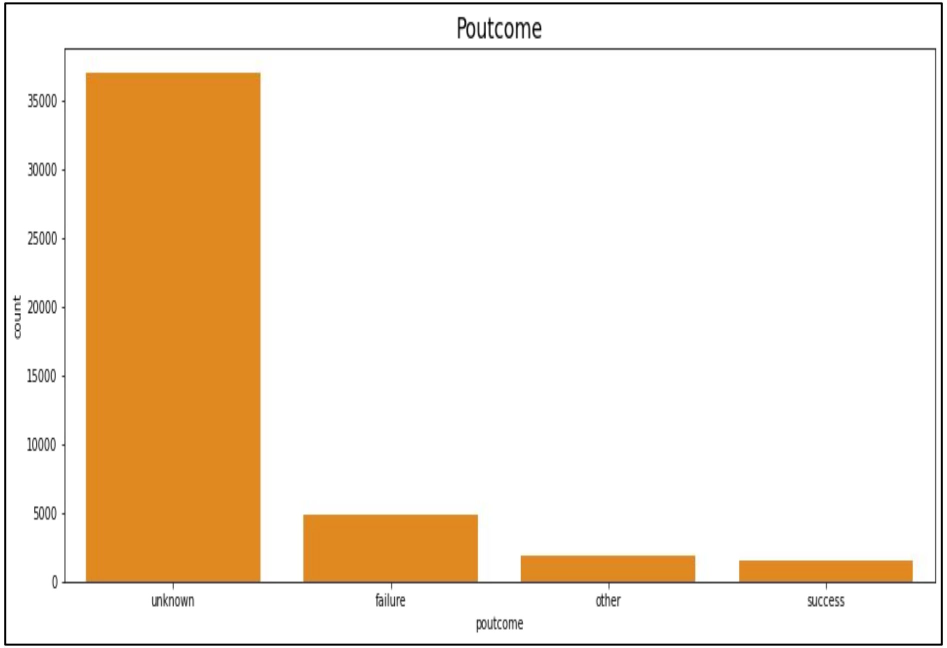
EDA(Continued...)



- Data in month of may is high and less in Dec.
- The month of the highest level of marketing activity was the month of May.

EDA(Continued...)

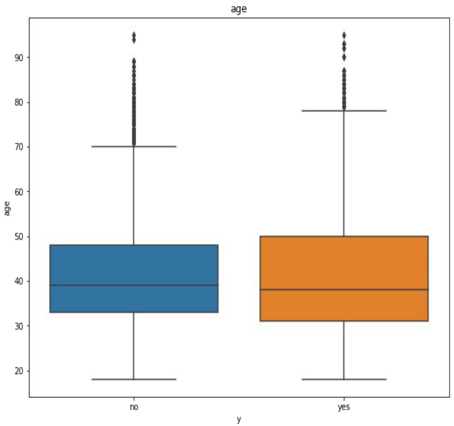


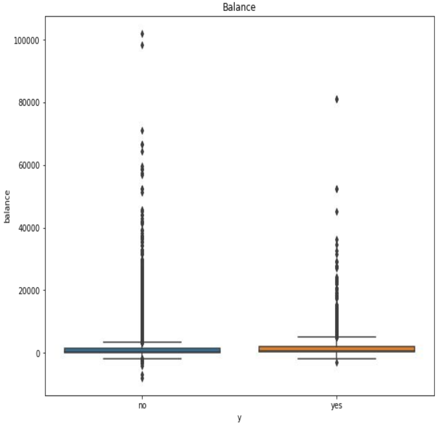


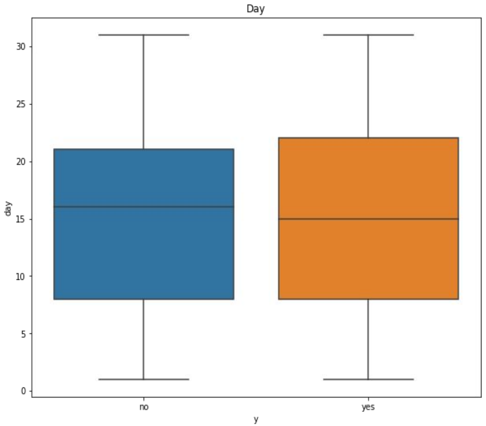
- Most people are contacted more in cellular than telephone.
- More people contacted on cellular by bank have subscribed the deposits.

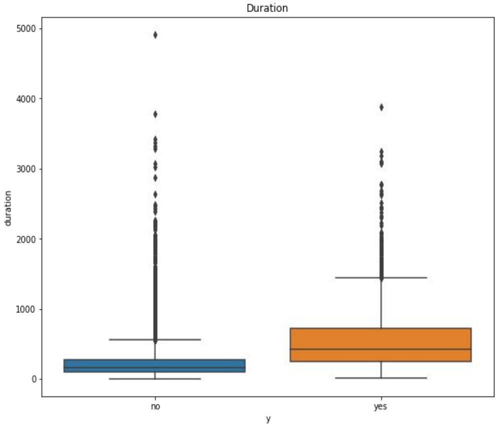
- Majority of the outcome of the previous campaign is Non-Existent.
- People whose previous outcome is non-existent have actually subscribed more.

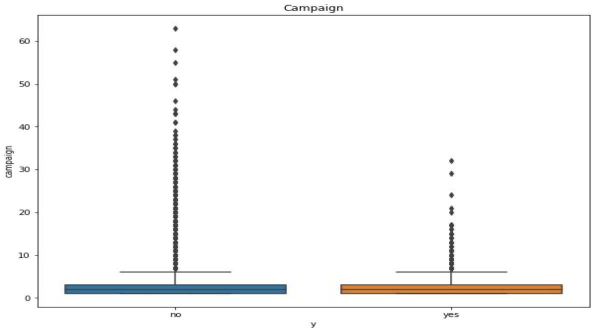
EDA(Continued...)

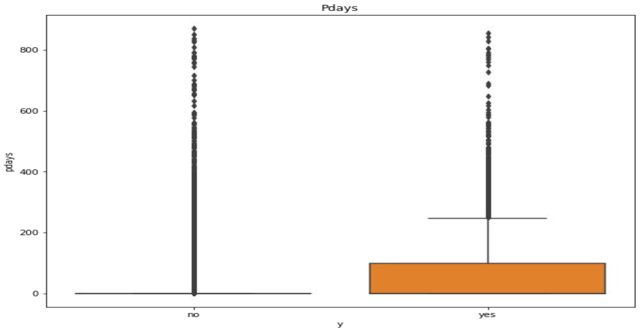

age

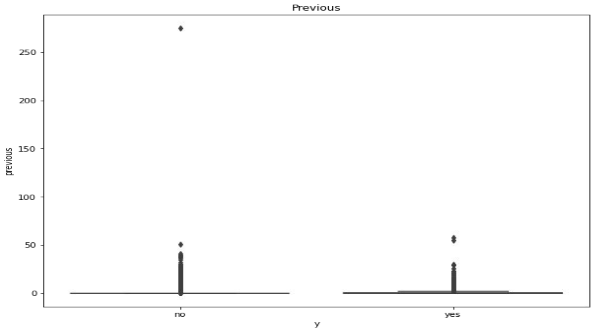

Balance


Day

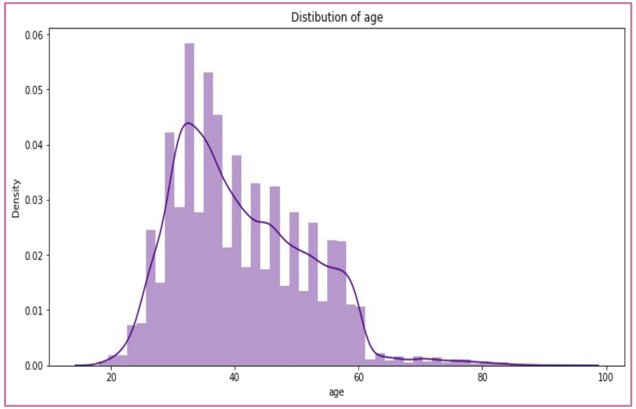

Duration

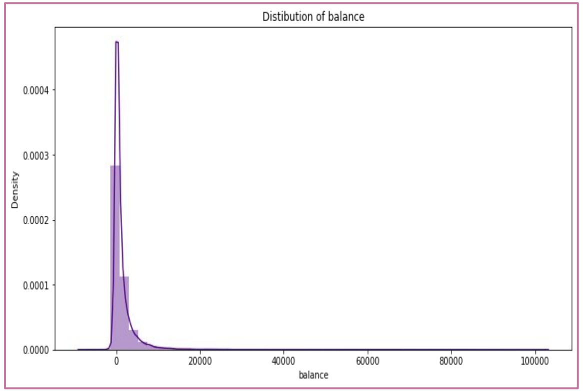

Campaign

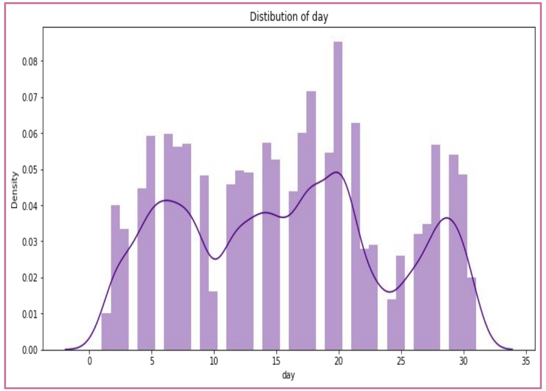

Pdays

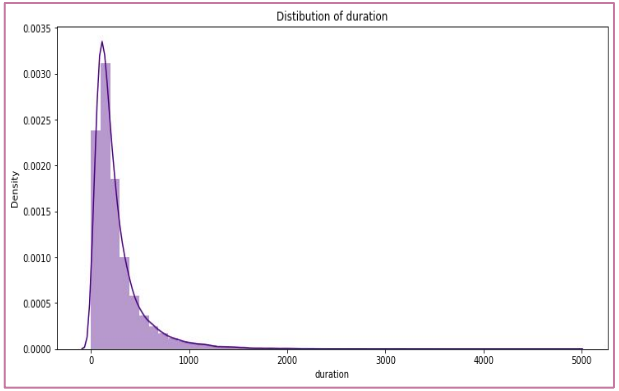

Previous

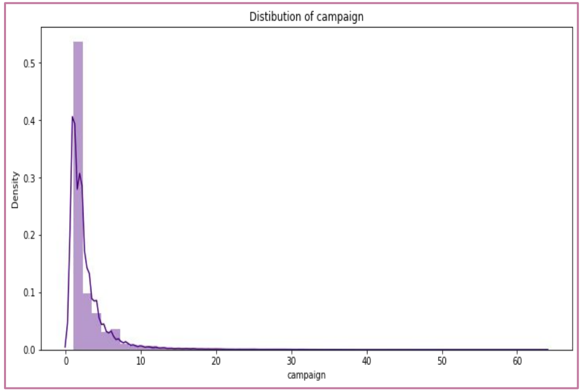
EDA(Continued..)

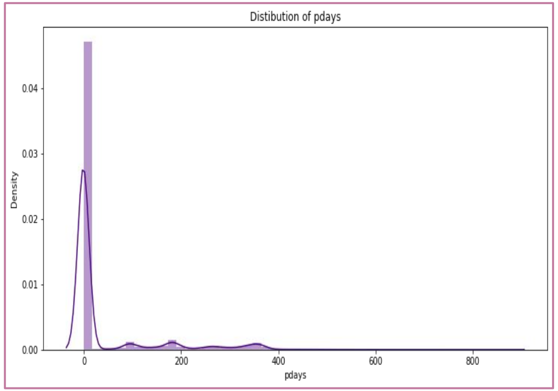




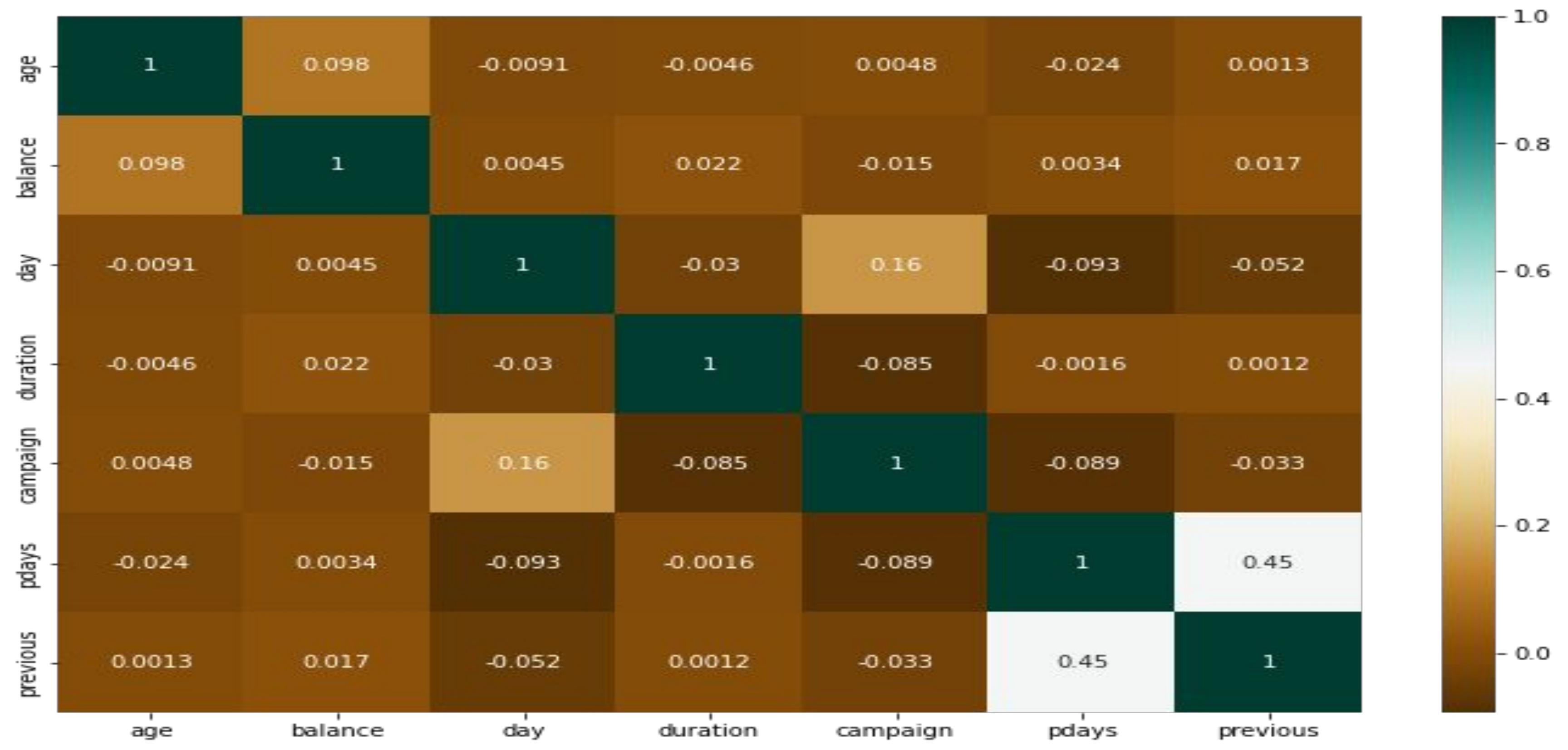








Correlation



	age	balance	day	duration	campaign	pdays	previous
age	1	0.098	-0.0091	-0.0046	0.0048	-0.024	0.0013
balance	0.098	1	0.0045	0.022	-0.015	0.0034	0.017
day	-0.0091	0.0045	1	-0.03	0.16	-0.093	-0.052
duration	-0.0046	0.022	-0.03	1	-0.085	-0.0016	0.0012
campaign	0.0048	-0.015	0.16	-0.085	1	-0.089	-0.033
pdays	-0.024	0.0034	-0.093	-0.0016	-0.089	1	0.45
previous	0.0013	0.017	-0.052	0.0012	-0.033	0.45	1

Model Implementation

Logistic Regression :-

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression is used for solving the classification problems.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Best Parameter :-

c : 0.1

ROC-AUC Score :-

Train Data	0.93
Test Data	0.92

Decision Tree :-

Decision trees are a way of modeling decisions and outcomes, mapping decisions in a branching structure. Decision trees are used to calculate the potential success of different series of decisions made to achieve a specific goal. The concept of a decision tree existed long before machine learning, as it can be used to manually model operational decisions like a flowchart. They are commonly taught and utilised in business, economics and operation management sectors as an approach to analysing organisational decision making.

Best Parameter :-

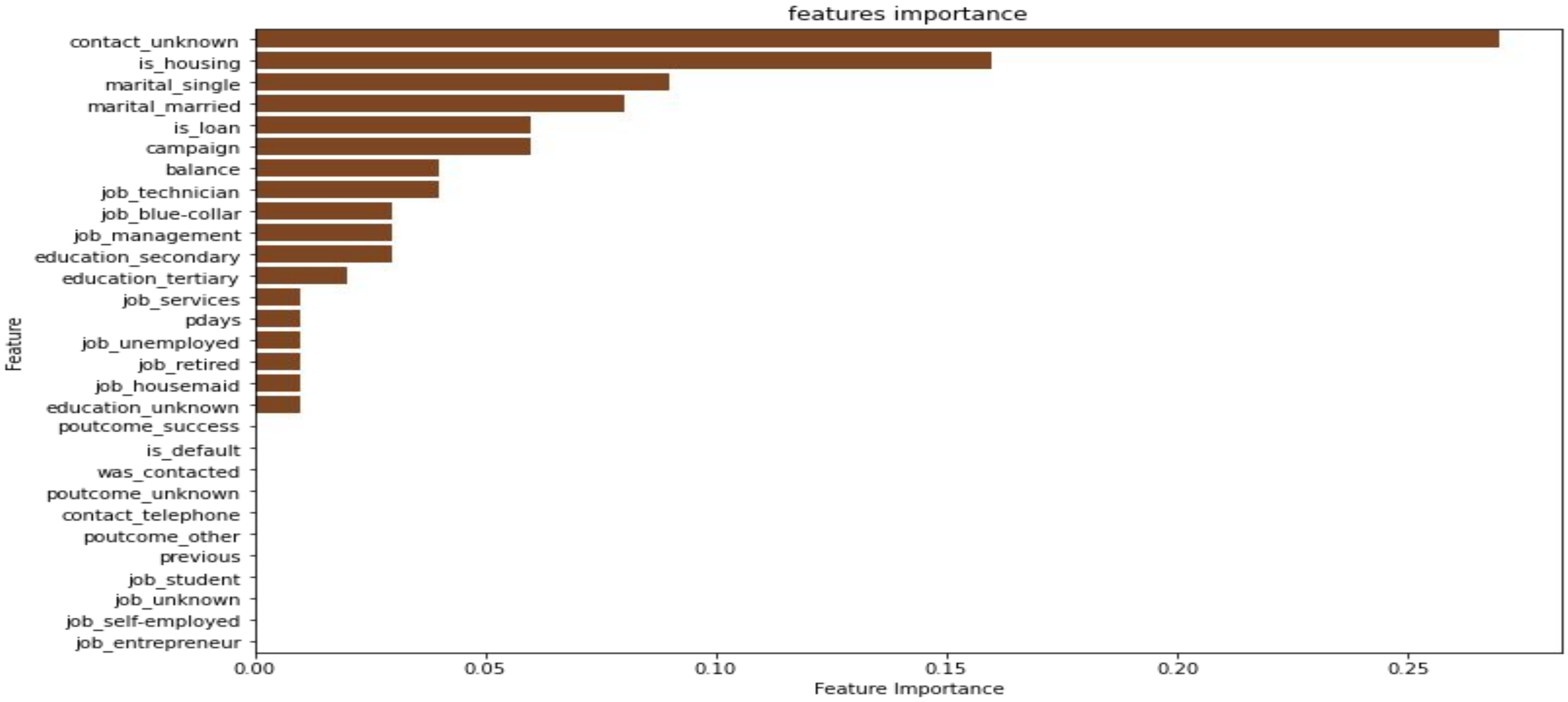
Mean_Sample_Leaf : 10
Max_Depth : 9
Mean_Sample_split : 20

ROC-AUC Score :-

Train Data	0.92
Test Data	0.90

Decision Tree :-

Decision Tree Features Importance



XGBoost Classifier :-

XGBoost classifier is a gradient boosting method that combines the regression tree . XGBoost is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created in sequential form.

Best Parameter :-

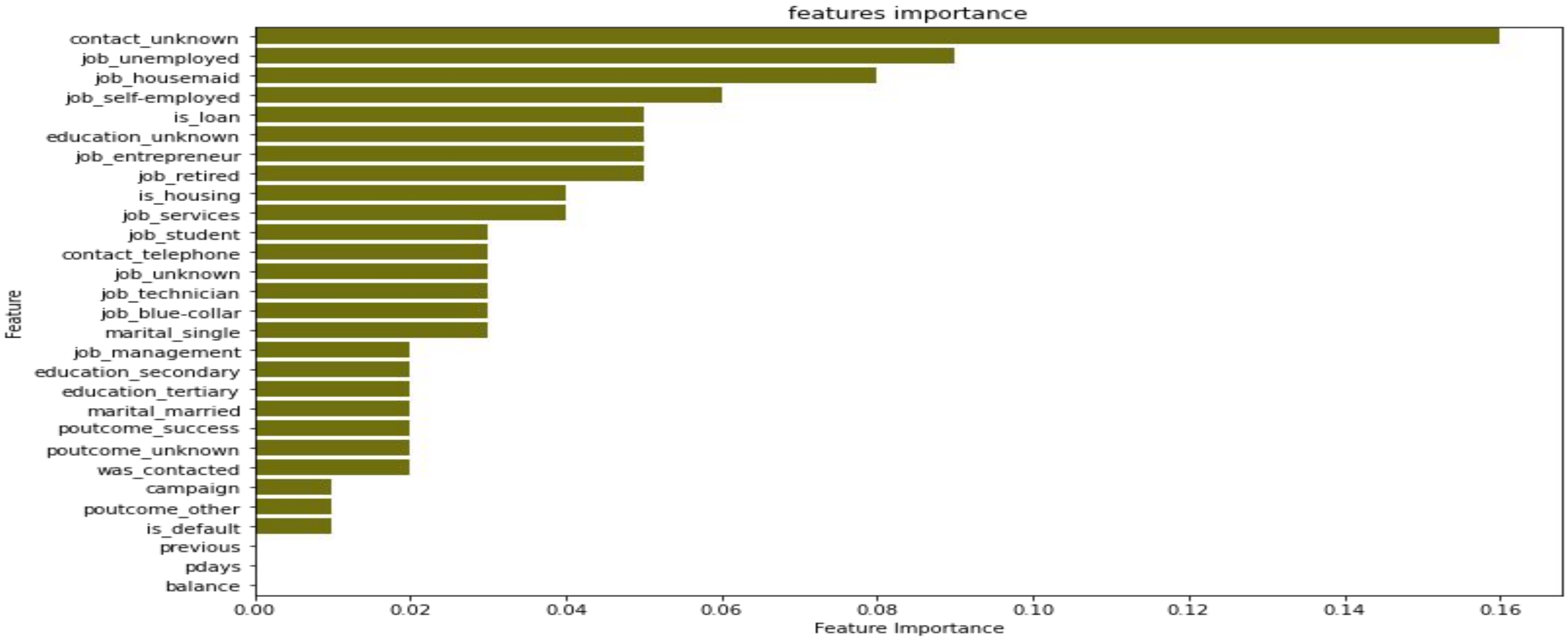
Learning rate : 0.5
Max_Depth : 9
N_estimators : 125

ROC-AUC Score :-

Train Data	0.95
Test Data	0.90

XGBoost Classifier :-

XGBoost Features Importance



K-Nearest Neighbor:-

K-Nearest Neighbor (KNN) Algorithm for Machine Learning K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

Best Parameter :-

n_neighbors : 27

ROC-AUC Score :-

Train Data	0.95
Test Data	0.93

Hyperparameter Tuning Evaluation

Model	Test AUC	Test Accuracy	F1-score	Precision
Logistic Regression	0.92	0.86	0.87	0.89
Decision Trees	0.90	0.83	0.84	0.85
XGBoost	0.93	0.90	0.91	0.92
K-NN	0.93	0.88	0.88	0.91

Challenges

- Handling Imbalanced Dataset
- Feature Engineering
- Optimising The Model

Conclusion

- For age, most of the customers are in the age range of 30-40.
- For balance, above 1000$ is like to subscribe a term deposit.
- The model can help to classify the customers on the basis on which they deposit or not
- The model helps to target the right customer rather than wasting time on wrong customer
- Comparing to all algorithms XGboost algorithm has best accuracy score and ROC-AUC score . So it is concluded as optimal model.

THANK YOU